Draft version September 19, 2006
Preprint typeset using LaTeX style emulateapj v. 11/26/03

MAPPING POPULATION SYNTHESIS EVENT RATES ON MODEL PARAMETERS II: CONVERGENCE AND ACCURACY OF MULTIDIMENSIONAL FITS

R. O'Shaughnessy, V. Kalogera, & K. Belczynski

Draft version September 19, 2006

ABSTRACT

To better understand several key event rates involving binary evolution and compact objects in Milky Way-like galaxies, we perform a methodical parameter study of the *StarTrack* population synthesis code. We significantly generalize earlier studies, and we provide thoroughly tested and well understood *multidimensional fits* for event rate results. These fits can be used in lieu of large-scale population calculations which are often forbiddingly computationally demanding. We anticipate that these efficient tools will facilitate the exploration of the dependence of rate predictions on a wide range binary evolution parameters and allow the derivation of constraints on these parameters, given empirical rate constraints and accounting for fitting errors.

Subject headings: binaries:close — stars:evolution — stars:neutron – black hole physics

1. INTRODUCTION

Models for binary stellar evolution and population syntheses are necessary to provide quantitative theoretical predictions for the relative likelihood of assorted events involving the evolution of binary stars. The resulting predictions are particularly critical when no empirical estimates exist for topics of immediate astrophysical interest, such as mergers of double-compact object (DCO) through the emission of gravitational waves. The most practical and widely applied binary population synthesis codes available in the community – such as the *StarTrack* code described in Belczynski et al. (2002) and significantly updated in Belczynski et al. (2006); the BSE code, described in Hurley et al. (2002); the SeBa code described in Portegies Zwart & Verbunt (1996); and the StarFaster code described in Fryer et al. (1998) – rely on a large set of fairly simple parameterized rules to characterize many complex and often ill-understood physical processes. Unfortunately, current binary population synthesis codes are greatly and often forbiddingly computationally demanding (depending on their level of sophistication): even with substantial simplifications, exploring the entire parameter space is beyond present-day computational capability.

However, observational information can provide us with constraints that help us improve our understanding of massive binary evolution. For example, pulsar searches continue to discover and refine observations of isolated pulsars and new binary pulsar systems; e.g., see Lorimer (2005) for a review. Specifically, the samples of binary pulsars with neutron star and relatively massive white dwarf companions have been used for a statistical derivation of empirical rate estimates for their formation (most recently see Kim et al. (2003), Kim et al. (2004), and Kalogera et al. (2005) and references therein. Additionally, many ground-based gravitational wave detectors now operating at or near design sensitivity (i.e., LIGO, GEO, TAMA) are designed to detect the late stages of double compact object (DCO) inspiral and merger. Based only on early-stage data, these instruments have already provided conservative upper limits to certain DCO merger rates (see,e.g. Abbott et al. 2005b,a). With LIGO now very close to design sensitivity, a year of LIGO data could definitively exclude (and even possibly confirm) the most extremely optimistic theoretical predictions for BH-BH merger rates (see, e.g. O'Shaughnessy et al. 2005a, for a range of BH-BH merger rates arising from binary evolution in Milky Way-like galaxies). Thus, gravitational-wave based upper limits (and, eventually, detections) will shortly provide constraints on theoretical models of DCO formation.

Faced with the availability of empirical rate constraints, and yet at first unable to quantitatively impose them on population synthesis predictions, O'Shaughnessy et al. (2005a) realized that any single unambiguous population synthesis prediction could be sampled loosely and then fit over the most sensitive population synthesis parameters. In the same study, we also presented a technique to accelerate the synthesis code used (*StarTrack*) to study a single target subpopulation, which we called "partitioning": we used experience gained from prior simulations to reject binary parameters highly unlikely to produce the current event of interest. O'Shaughnessy et al. (2005b) first applied these early fits to allow a direct comparison between *StarTrack*-produced population synthesis predictions and the observed formation rate for NS-NS binaries. Though only a small fraction (2%) of *StarTrack* models appeared consistent with the constraints, conceptual challenges with seven-dimensional visualization prevented O'Shaughnessy et al. (2005a) from clearly describing the constraint-satisfying region. A forthcoming paper, O'Shaughnessy et al. (2006c) will significantly extend this preliminary analysis, adding significantly more observational constraints as well as a clearer investigation of the constraint-satisfying models.

In this paper, we significantly generalize the analysis of O'Shaughnessy et al. (2005a), and we present a thorough discussion of our much-updated and vastly larger population synthesis archive and particularly of the fitting methods we employ to extract predictions. In the present analysis we consider a critically notable new element: we explicitly employ realistic, yet simple, statistical problems (we denote them *toy models*) to demonstrate how our fitting procedure can diagnose and avoid several obstacles toward accurate fits and fit applica-

tions. For example, since our high-dimensional polynomial fits (even with relatively low polynomial order) commonly involve roughly half as many points as parameters, we develop a simple self-consistency diagnostic to verify overfitting does not occur. Additionally, since we wish to reconstruct constraint-satisfying regions using these fits, we study the degree to which a toy-model constraint-satisfying region agrees with the expected volume. We next apply these statistical lessons to a larger (both in number and parameter space covered), much more *heterogeneous* and *irregularly sampled* archive of population synthesis calculations than that used in O'Shaughnessy et al. (2005a). We extract our best estimates for a range of DCO and supernovae event rates for Milky Way-like star forming conditions. Moreover, we estimate the reliability of these rates, taking into account possible systematic errors associated in their estimation. Finally, in a purely technical detail, we use new fitting basis polynomials, as described in Appendix A.

2. POPULATION SYNTHESIS ARCHIVES

Population synthesis simulations can be extremely computationally demanding: even though *StarTrack* can fully evolve roughly 10^3 binaries of interest[1] per CPU-hour with modern-day processors, because some double compact objects form very infrequently (e.g., black holes, which occur roughly once every $\approx 10^{-4}$ binaries evolved), a representative sample of stellar systems often contains $10^{4.5}-10^{6.5}$ binaries and requires hundreds of CPU hours to complete. Additionally, since population synthesis rate predictions depend delicately on model parameters, the *computation time needed* to build up a sufficiently representative collection of stellar systems – one where some event of interest occurs many times – varies considerably depending on astrophysical assumptions. Given the prohibitive computational demands of a brute-force approach, we took advantage of several simplifications originally developed in O'Shaughnessy et al. (2005a) to assemble our archive of roughly 3000 population synthesis simulations, upon which our fits our predicated. In this section we briefly describe how those archives were generated and how we identify and extract event rates for several processes of interest.

2.1. StarTrack *population synthesis code*

We estimate formation and merger rates for several classes of double compact objects using the *StarTrack* code first developed by Belczynski et al. (2002) [hereafter BKB] and recently significantly updated and tested as described in detail in Belczynski et al. (2006). Like other population synthesis codes, *StarTrack* evolves binaries from their birth (drawn randomly from specified birth distributions) to the present, tracking the stellar and binary parameters. For any class of events that is *identifiable within the code*, such as supernovae or DCO mergers, we estimate event rates by: (i) taking the average event rate within the simulation (i.e., the total number of events seen within some simulation divided by the duration of that simulation), and (ii) renormalizing by a scale factor that depends on properties of the simulation (i.e., the number of binaries simulated and the binary birth distributions assumed) and of the Milky Way as a whole (i.e., the present-day star formation rate): $R = s \times n/T$, where n is the number of events observed and T is assumed $T = 10$ Gyr. The scale factor s is a ratio $s = N_g/N_{eff}$ between the number of stellar systems in the Milky Way and the number of stellar systems we have effectively sampled to select our n events of interest. The effective sample size N_{eff} is the number of stellar systems needed, on average, to produce N stellar systems with $m_1 > 4M_\odot$, if all systems are drawn from an IMF which extends from the hydrogen burning limit $m = 0.08M_\odot$ to $m = 150_\odot$:

$$N_{eff} = N/\int_{4M_\odot}^{150M_\odot} dm\ \phi(m)\ . \quad (1)$$

We use a Kroupa IMF: $\phi(m) \propto m^{-1.3}$ if $m \in [0.08, 0.5]M_\odot$, $\propto m^{-2.2}$ if $m \in [0.5, 1]M_\odot$, and $\propto m^{-2.7}$ if $m > 1M_\odot$. Finally, we choose N_g so that N_g times the average mass $\langle m_{tot}\rangle$ (according to our IMFs for m_1 and m_2/m_1, and according to the assumed fraction f_b of stellar *systems* which are binary) of each stellar system is equal to the total mass which should be formed in stars over the $T = 10$Gyr lifetime of the Milky Way, $\dot{M}T$:

$$N_g = \frac{\dot{M}T}{\langle m_{tot}\rangle} = \frac{\dot{M}T}{\langle m_1\rangle\,(1 + f_b\,\langle m_2/m_1\rangle)}\ . \quad (2)$$

For the rate $\dot{M}$ at which mass is born in stars, we use the empirical estimate $\dot{M} \approx 3.5 M_\odot \mathrm{yr}^{-1}$ (Rana 1991; Lacey & Fall 1985). The average values of m_1 and m_2/m_1 are found from the Kroupa IMF and the assumed distribution for m_2/m_1 presented in O'Shaughnessy et al. (2005a).

Our approach gives only the *average* event rate. The present-day merger rate agrees with this quantity when most mergers occur relatively promptly after their birth (relative to the age of the Milky Way and hence duration of the star formation phase, e.g., < 100 Myr). Some DCOs – notably double BH binaries – have substantial delays between birth and merger, introducing a strong time dependence to the merger rate. The technique described above will significantly *underestimate* these rates. This point will be addressed in considerably more detail, both for the Milky Way and for a heterogeneous galaxy population in a forthcoming paper by O'Shaughnessy et al. (2006a) and in greater detail in O'Shaughnessy et al. (2006b).

2.2. *Parameters varied in archives*

Our multi-year experience in modeling of binary compact objects with *StarTrack* clearly indicates that there are seven parameters that strongly influence compact object merger rates (see e.g., Belczynski et al. (2002)): the supernova kick distribution (3 parameters $\sigma_{1,2}$ and s describing a superposition of two independent maxwellians), the massive stellar wind strength w (1), the common-envelope energy transfer efficiency $\alpha\lambda$ (1), the fraction of mass accreted by the accretor in phases of non-conservative mass transfer f_a (1), and the binary mass ratio distribution, as described by a negative power-law index r (1). We allow the dimensionless parameters $\alpha\lambda, f_a$, w and s to run from 0 to 1; the dimensionless r can be between 0 and 3; and finally we vary the dispersion of either component of a bimodal Maxwellian σ_1, σ_2 from 0 to 1000 km/s.

[1] Specifically $m_1 > 4$; see the discussion below.

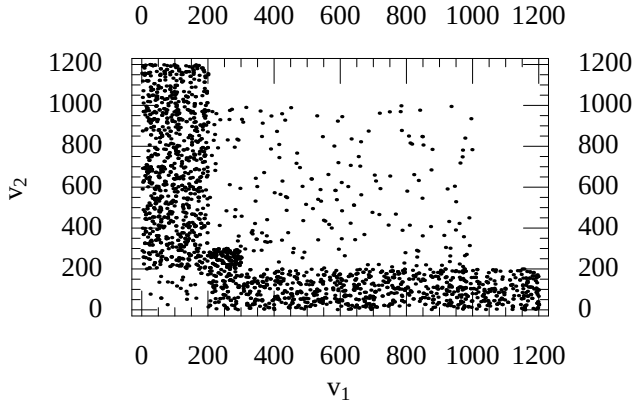

FIG. 1.— Scatter plot of the two bimodal kick velocity dispersions σ_1, σ_2 for the population synthesis archives used to evaluate the merger rate of NS-NS binaries [denoted NSNS(m)]. The strong bias is introduced by incorporating high-density and computationally expensive simulations from O'Shaughnessy et al. (2005a) and O'Shaughnessy et al. (2005b).

To improve our statistics, we combine results from several different databases of simulations. The most extensive database samples $\sigma_1 \in [200, 1200]$ and $\sigma_2 \in [0, 200]$ very densely and was developed by O'Shaughnessy et al. (2005a) and O'Shaughnessy et al. (2005b). A second archive, significantly less dense due to computational resource limitations, allows both dispersions to run uniformly from 0 to 1000 km/s. [Additional archives include, for example, a set chosen to better-sample kick parameters that best correspond to observations of pulsar proper motions Arzoumanian et al. (1999); Hobbs et al. (2005).] Consequently, as shown in Figure 1, our archived results do not uniformly sample the kick-related parameters through this range. This undersampling likely plays a role in the relatively poor *global* convergence of fits for physical parameters, as described in § 4. Nonetheless, our sampling fairly thoroughly explores the most physically likely regimes suggested by Hobbs et al. (2005) and Arzoumanian et al. (1999).

2.3. *Event identification*

Most events of physical interest are uniquely and fairly unambiguously identifiable within the code. Type II and Ib/c supernovae events are distinguished by the presence or absence of a hydrogen-rich envelope at the supernova event. We also record DCOs which merge, so we can unambiguously determine the number of BH-BH, BH-NS, NS-NS, and WD-NS merger events which occur in a simulation. [These event classes will be denoted BHBH(m), BHNS(m), NSNS(m), and WDNS(m) for brevity henceforth.] Since the code also tracks binary eccentricity, for example, we can identify those WD-NS binaries which end their evolution with a non-zero eccentricy [denoted WDNS(e)].

When constructing our archived population synthesis results, we do not record detailed information about the nature and amount of any mass transfer onto the first-born NS. We therefore cannot determine the degree to which pulsars are recycled. However, we do record whether any mass transfer occurs. Thus for the purposes of identifying a class of potentially recycled ("visible") wide NS-NS binaries [denoted NSNS(vw)], we assume any system undergoing *non-CE* mass transfer recycles its NS primary.

2.4. *Practical Archive Generation and Resolution, with Partitions and Heterogeneous Targets*

The accuracy to which each population synthesis event rate prediction is known, $\approx 1/\sqrt{n}$, is uniquely set by the number of events n seen within that simulation. For this reason, O'Shaughnessy et al. (2005a) (i) designed their population synthesis runs to continue until a *fixed number of events were seen* and (ii) used results only from such targeted simulations, where a minimum number of events was guaranteed. In contrast in this study, given limited computational resources and a wide range of targets for which predictions are needed, we extract *all possible information* from each simulation: whenever possible, we make an estimate of each event rate of interest.

However, it is important to note that most of our simulations employ some degree of accelerating simplification which can *bias* estimates. To give the most extreme example: the most accurate estimates for the BH-BH merger rate come from population synthesis runs which evolve only a subset of possible progenitor binaries by using *partitions*. This subset has been shown to include the progenitors of the vast majority of BH-BH binaries, but very few progenitors of NS-NS binaries and other less massive DCOs (for more information, see O'Shaughnessy et al. 2005a, ; we continue to employ the same partition they devised). Similarly, the vast majority of simulations used to study NS-NS event rates (i) use a similar partition to reduce contamination from white dwarf binaries and (ii) terminate the evolution of any binary immediately after any WD forms. These strong biases make data from these two types of simulations inappropriate for use in, for example, estimating the WD-NS merger rate. For this reason, the number of population synthesis archives N available to make predictions varies significantly across the various target event types; see Table 1.

Furthermore, the accuracy each simulation can provide varies considerably: unlike earlier studies, the number of event samples n in each archive is not guaranteed to be greater than a minimum value. However, usually we have at least one event for each event type: N_{s+}, the number of unbiased population synthesis results containing one or more events is usually very close to N_s, the total number of unbiased population synthesis simulations available. And in most cases a significant proportion of our sample contains enough events to determine the rate to better than 30% (i.e., $n \geq 10$).

3. FITTING: TOY MODELS

Given the prohibitive computational demands of direct population synthesis simulation described in § 2, we use fits based on archived results of population synthesis runs as a surrogate for repeated detailed simulations. Confidence in our results is therefore tied intimately to confidence in the quality of these fits. However, even low order fits in seven dimensions involve many parameters: to fit any nontrivial function, we *must* fit roughly a handful of data points per parameter. To build confidence, we want to show that fit order chosen adequately describes the data without overfitting. More delicately, we also

want to suggest that the key end product of our calculation, the "constraint-satisfying region", does not depend sensitively on the fit details or on random accidents in the data (i.e., any different monte carlo realization of our simulations should yield the same result). To test our fitting methods within a well-controlled and -defined sandbox pipeline, we apply a simplified version of our procedure to a suite of known test functions.

3.1. *Fitting procedure*

This section describes this simplified process as well as the features of the true fitting problem we seek to mimic.

First, to mimic our set of randomly-chosen population synthesis simulation parameters, we pick a set $\vec{x}_{\alpha=1\ldots N_s}$ of N_s randomly chosen sample points in the 7-dimensional cube $\{\vec{x}_\alpha | x_\alpha^{k=1\ldots 7} \in [0,1]\}$. Next, to mimic the results of performing each population synthesis simulation consisting of N progenitor binaries, for each point $\vec{x}_\alpha$, we generate a poisson-distributed integer n_α with mean $\mu_\alpha = \mu(\vec{x}_\alpha)$. [The toy-model quantity μ_α/N corresponds in population synthesis studies to the small probability that any progenitor binary is, say, a NS-NS merger; for a finite fixed number of progenitor binaries N, the *observed* sample of merging binaries n_α would be poisson-distributed, with mean μ_α.] Since we intend to fit to the *logarithm* of the rate, we define $M(x)$ so $\mu(x) \equiv 10^{M(\vec{x}_\alpha)}$. Almost in exact correspondance to our goal of determining the number of events that occur within the lifetime of the Milky Way, our goal in these toy problems is to recover $\mu(\vec{x})$, or equivalently $M(\vec{x})$, throughout the 7-dimensional cube. We represent our best-fit estimator for these two quantities by $\hat{M}(\vec{x})$ and $\hat{\mu}(\vec{x})$. [Hereafter, we will often omit the dependence on $\vec{x}$, which is always implied.] Further, we will expand $\hat{M}(\vec{x})$ in terms of basis functions $\{T_A(\vec{x}) | A = 1 \ldots N_q\}$. For this paper, we will use T_A to be the set of all seven-dimensional basis polynomials of order $\leq q$, which consists of the $N_q = (7+q-1)!/6!q!$ distinct polymomials of the form $x_1^{a_1} x_2^{b_2} \ldots x_7^{a_7}$ with $\sum_k a_k \leq q$.

Since the sampling error at each point follows precisely a Poisson distribution, the maximum likelihood polynomial $\hat{M}_q$ of order $\leq q$ can be unambiguously determined by maximizing the log-likelihood function for observations n_α drawn from poisson parameters $\tilde{\mu}$

$$L(\tilde{\mu}) = \ln \prod_\alpha \frac{\tilde{\mu}^{n_\alpha}}{n_\alpha!} e^{-\tilde{\mu}} \tag{3}$$

$$L(\tilde{\mu}) \propto \frac{1}{N_s - N_q} \sum_\alpha \left[n_\alpha - \tilde{\mu}(x_\alpha) + n_\alpha \ln\left(\frac{\tilde{\mu}(x_\alpha)}{n_\alpha}\right)\right] + \text{constant} \tag{4}$$

In the maximization, we fix the observed data n_α and vary over all possible functions $\tilde{\mu}$, or equivalently over the N_s-dimensional vector $\tilde{\mu}(x_\alpha)$. By definition, the values $\hat{\mu}(x_\alpha)$ which maximizes the log-likelihood provides the maximum-likelihood estimate of μ at the data points x_α. To better diagnose the degree to which the function $\hat{\mu}$ maximizes the log-likelihood, we define L_q to be

$$L_q \equiv \frac{1}{N_s - N_q} \sum_\alpha \left[n_\alpha - \hat{\mu}_q(\vec{x}_\alpha) + n_\alpha \ln\left(\frac{\hat{\mu}_q(\vec{x}_\alpha)}{n_\alpha}\right)\right] \tag{5}$$

[The proportionality and constant do not affect maximization of the log-likelihood.] However, as always in high-dimensional maximum-likelihood fits, the numerical maximization process need not converge on the global minimum. As a robust and deterministic alternative, we typically employ a weighted least-squares fit for most fitting purposes.

The error-weighted least-squares fit minimizes an weighted difference χ^2 beween our fit and the data, defined per degree of freedom:

$$\chi_q^2 \equiv \frac{1}{N_s - N_q} \sum_\alpha \left|\log_{10}\left[n_\alpha/\hat{\mu}_q(\vec{x}_\alpha)\right]\right|^2 n_\alpha \ln(10)^2 \tag{6}$$

$$= \frac{1}{N_s - N_q} \sum_\alpha |\log_{10} n_\alpha - \hat{M}_q(x_\alpha)|^2 / \Sigma_\alpha^2$$

where $\Sigma \equiv 1/\sqrt{n_\alpha}\ln(10)$ is a reasonable a priori estimate, based on the relative uncertainty (i.e., $\mu/\sqrt{\mu}$) of the poisson distribution, of the uncertainty in our estimate of $\log_{10}\mu(x_\alpha)$ based on the observed sample n_α. [Points with $n_\alpha = 0$ do not contribute to this sum and are ignored.] This least-squares approach has the distinct advantage of expediency (the coefficients for any $\hat{M}_q$ can be quickly determined from a handful of matrix operations) and familiarity. Additionally, when the n_α are large, poisson distributions converge to gaussians, implying that when we fit to well-resolved rates both the least-squares and maximum-likelihood maximization methods should recover the same result.

3.2. *Fit Diagnostics: Point Error*

In this section we describe a large collection of diagnostic procedures to estimate the error in using $\hat{M}_q$ to model M.

Comparing with a known solution: In this toy model, the function M being fitted is known and can be compared directly with the fit. The *global* rms deviation between it and our best-fit estimator $\hat{M}_q$ based on q-th order polynomials:

$$I_q = \left[\int d^7x |M - \hat{M}_q|^2\right]^{1/2} . \tag{7}$$

For the purposes of this paper, an optimal fit minimizes this difference I_q. This comparison cannot be made when the physical solution is not known, as when fitting to archived population synthesis data. However, this ideal diagnostic still provides a ruler against which other, fit-based diagnostics can be compared.

For a perfect fit, where $\hat{M}_q = M$, I_q should be zero. In practice, M need not lie within the span of our basis polynomials at low order q. We therefore define an auxiliary diagnostic, $\tilde{I}_q = \max_{\hat{M}_q} I_q$, to measure the minimum possible value that I_q can take over all possible functions in the span of $\{T_A\}$.

Consistency with data: When the true solution is not known, diagnostics can involve only the data and fits. Perhaps the simplest diagnostic is the rms deviation (per degree of freedom) between the observed values and our best-fit data:

$$\sigma_q^2 \equiv \frac{1}{N_s - N_q} \sum_\alpha (\log n_\alpha - \hat{M}_q(x_\alpha))^2 , \tag{8}$$

Given a perfect fit but allowing for sampling fluctions, this rms deviation on average should agree with the statistically expected sampling fluctuations: briefly, σ_q can

statistically at best be expected to reach $\sigma_{q,E}$,

$$\sigma_{q,E}^2 \equiv \sum_\alpha \frac{\Sigma_\alpha^2}{N_s - N_q} = \sum_\alpha [(N_s - N_q) n_\alpha (\ln 10)^2]^{-1} \quad (9)$$

In practice even the best possible fit allowed cannot perfectly represent the trial function, introducing errors characterized by $\tilde{I}_q$ [Eq. (7)]. When these systematic errors are large compared to statistical errors (i.e., $I_q \gg \sigma_{q,E}$), then we expect large rms deviation between the data and fit (σ_q): roughly speaking

$$\sigma_q^2 \approx I_q^2 + \sigma_{q,E}^2 \geq \tilde{I}_q^2 \; . \quad (10)$$

Statistical consistency with data: We can compare the degree to which the fits reproduce the data in a statistically plausible manner by evaluating the value of the log-likelihood L_q for the fit [Eq. (5)]. Less sensitively, we can compare the *nonzero* measurements with the fit using χ_q^2 [Eq. (6)]. Because of sampling fluctuations, these quantities will rarely be precisely zero: even when we know the function being fitted (i.e., $\hat{M} = M$), we expect L_q and χ_q^2 to average -0.5 and 1, respectively.

Large values for these two statistical diagnostics unambiguously indicate a "poor" fit, compared to the best possible fit that could be expected given sampling uncertainties associated with the data (n_α). However, even with large fit order q large values for these diagnostics are often unavoidable: since M need not lie within the span of $\{T_A\}$, the *magnitude* of these "statistical" diagnostics need not reflect the probability that the data and fit are consistent. Likewise, *small* values of these two statistical diagnostics do not guarantee a good fit: "statistically plausible" fits exist with enough parameters to match most of the observed data yet which violently fail elsewhere (as measured by I_q); see Figure 6 for examples.

Blind test: Philosophically, the best test of a fits' performance is in a *blind test*, where the fit $\hat{M}_q$ is compared against a set B of N_t test points (x_β, n_β) completely distinct from the set A of N_s points on which the fit was developed. Specifically, we can evaluate the analagous χ^2 quantity per degree of freedom,

$$\chi_{q,b}^2 \equiv \frac{1}{N_t} \sum_\beta \left| \log_{10} \left[n_\beta / \hat{\mu}_q(x_\beta) \right] \right|^2 n_\beta \ln(10)^2 \; . \quad (11)$$

Blind tests diagnose whether, in the rush to minimize the difference between the *observed* data and our trial function $\hat{M}_q$, too many parameters have been introduced, allowing for spurious oscillations between data points. Ideally, the differences between the blind-test data B and the fit are statistically similar to the corresponding differences with the fitted data, so ideally $\chi_{q,b}^2 \approx \chi_q^2$.

For the purposes of this paper, to insure the test sample covers a reasonable proportion of the data sample we require $N_t = 0.1 N_s$.

Comparison of different fit orders: Finally, we can test the degree to which the fit converges by comparing the rms difference between fits of successive orders:

$$J_q = \left[\int d^7x |\hat{M}_q - \hat{M}_{q-1}|^2 \right]^{1/2} \; . \quad (12)$$

Since adjoining increasingly many parameters to fit the data produces strong spurious oscillations between data points, this integral qualitatively diagnoses *overfitting*.

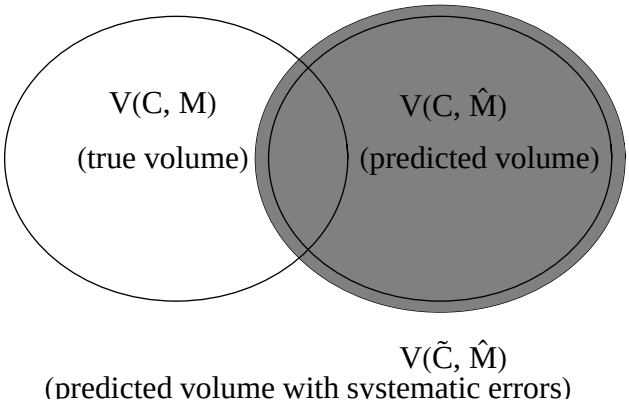

FIG. 2.— Illustration of the three types of constraint-satisfying volumes discussed in the paper $V(C, M)$, $V(C, \hat{M})$, and $V(\tilde{(C)}, \hat{M})$. Because of limitations in our population synthesis sample (both each simulation's resolution, as well as in the number and placement of simulations in the space of population synthesis models), the volume of population synthesis models which *actually* satisfies some set of observational constraints $V(C, M)$ does not necessarily correspond to the volume *predicted* $V(C, \hat{M})$, even when the constraint interval is increased by an estimate of the systematic errors $V(\tilde{C}, \hat{M})$. The text defines diagnostics $r_\pm$ [Eq. (**??**)] to characterize the degree to which any two volumes agree.

3.3. *Fit diagnostics: Region Error*

With an excellent fit, the predicted constraint-satisfying regions (from the fit) should agree closely to the true constraint-satisfying regions (from the model). With a less-accurate fit, systematic errors must be incorporated, so the predicted volume reliably contains all models which satisfy the true constraints. To quantify the comparison, we present several diagnostics which for the differences between predicted and true constraint-satisfying volumes.

Given a logarithmic "rate constraint" interval $C = \{c_1, c_2 | c_1 < c_2\}$ and "rate function" $\mu = 10^M$, a constraint-satisfying subset volume $V(C, M)$ of the seven-dimensional unit cube exists such that, everywhere within that region, $c_1 \leq M \leq c_2$. A more conservative constraint interval $\tilde{C}(\epsilon) \equiv \{c_1 - \epsilon, c_2 + \epsilon\}$, the original constraint interval broadened by a tolerance x, produces a larger constraint-satisfying volume and a conservative overestimate of the constraint-satisfying region.

Our statistical estimator $\hat{M}_q$ for M produces a different region $V(C, \hat{M}_q)$; see Figure 2. We characterize the difference between the two by (i) the fraction r_+ of constraint-satisfying points correctly identified by the predicted ("hatted") constraint-satisfying region and (ii) the fraction r_- of points correctly predicted to be constraint satisfying:

$$r_+ \equiv \frac{|V(C, \hat{M}) \cap V(C, M)|}{|V(C, M)|} \quad (13)$$

$$r_- \equiv \frac{|V(C, \hat{M}) \cap V(C, M)|}{|V(C, \hat{M})|} \quad (14)$$

$$(15)$$

where $|V|$ denotes the volume of the set V. The ratio of these quantities, $r_-/r_+ = |V(C, \hat{M})|/|V(C, M)|$, indicates the accuracy to which our estimator $\hat{M}$ pre-

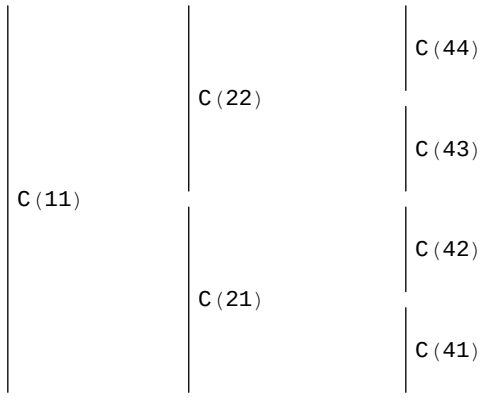

FIG. 3.— Illustration of how the nested constraint subintervals C_{pz} cover the resolvable range of M.

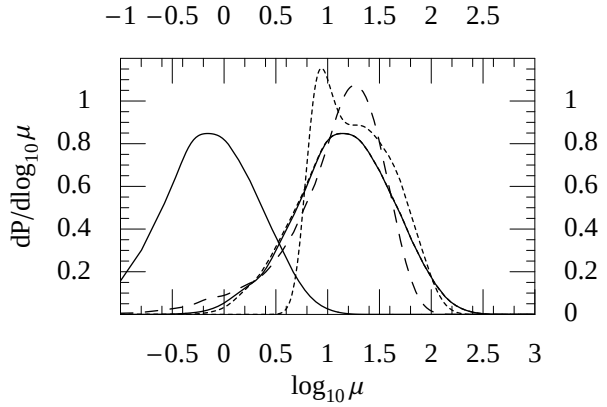

FIG. 4.— Relative frequency for different values of μ, expressed as a probability distribution for the random variable μ. The three nearly superposed curves (solid, dotted, and dashed) correspond to test functions 1,3, and 4 respectively. The solid curve with mean near zero corresponds to the trial function $\mu_2 = \mu_1/20$.

dicts the fraction of constraint-satisfying models. To be conservative, a predicted constraint-satisfying region should include as much of the true constraint-satisfying region as possible, or have r_+ close to 1; conversely, it should include as little as possible of superfluous points, or $1 - r_- \approx 0$.

Strong systematic errors, as when the target function μ varies in a way our basis functions cannot easily mimic, can seriously limit fit accuracy and in turn our ability to reliably predict the constraint-satisfying region. To extend the constraint-satisfying interval to better encompass the true constraint-satisfying volume $V(C, M)$, we introduce the tolerance-broadened intervals $\tilde{C}$, which are precisely C broadened by a tolerance J_q. As will be demonstrated below, since J_q approximates the rms point error in the fit, any population synthesis model which *actually* satisfies the constraints very likely has been predicted by the fit to lie within this constraint-broadened interval, and thus lies within $V(\tilde{C}, \hat{M}_q)$ (i.e., it has r_+ near 1). However, by enlargening the predicted volume we include substantially more superfluous points (i.e., r_- often lies considerably below 1).

The fit also depends sensitively on (i) the length and (ii) the placement of the constraint interval in comparison to the range of M. Very narrow constraints are difficult to satisfy and more susceptible to error, whereas constraints placed near the limits of M depend sensitively on accurate modeling of M near extrema and boundaries (particularly for intervals near the poorly-sampled *lower* limits of M). Rather than consider *all possible* constraint intervals, we introduce a fairly generic set of "basis" constraint subintervals C_{pz} and corresponding region diagnostics $r_{\pm,pz}$, where p is the number of these nonoverlapping equal-length subintervals and z is an index from 1 to p; see Figure 3. For each p, the nonoverlapping equal-length intervals $C_{p,z}$ cover the full range of M, from its maximum down to either zero or its minimum, whichever comes first. As p increases, the constraint subintervals grow ever narrower; our ability to correctly predict the constraint-satisfying volumes $V(C_{pz}, M)$ grows lower.

3.4. *Trial functions*

Ideally, our fitting procedure will recover all functions $\mu(x)$ which roughly correspond to those seen in population synthesis simulations. [By analogy, our fitting procedure will allow us to best estimate how *StarTrack*'s model predictions for event rates $\mathcal{R}$ depend on its input parameters.] To mimic the event counts usually seen in our population synthesis simulations, we choose the following suite of test functions for the "unknown prediction" function $\mu(x)$:

$$\mu_1(x) = 10^{4[-\vec{x}\cdot\vec{x}/7+1]-3/2} \quad (16)$$

$$\mu_2(x) = \mu_1(x)/20 \quad (17)$$

$$\mu_3(x) = \mu_1(x) 10^{2[\tanh(1-4(x-1)\cdot(x-1))+\tanh(27)]} \quad (18)$$

$$\mu_4(x) = 10^{2(1-\vec{x}\cdot\vec{x}/7-v_5 v_6 v_7)} \quad (19)$$

$$\mu_5(x) = 2\left(8 - 1.5\vec{x}\cdot\vec{x} + 8(\vec{x}\cdot\vec{x} - 3)^2\right) \quad (20)$$

As shown in Figure 4, all except μ_2 have been chosen to cover nearly an identical range with roughly the same probability of taking on any value. Further, they are designed to give poor but measurable estimates for $\mu(x)$: most have $\mu(x)$ on average around 15 with a standard deviation in $M(x)$ of 0.4.

3.5. *Application: Choosing an Optimal Fit based on Multiple Diagnostics*

Good convergence towards a consistent result is easily identified from the consistent behavior of our various fit diagnostics. As an example, we consider our fitting scheme applied to a Monte Carlo sampling of $\mu_1(x)$. Specifically, for each of set of trial sample sizes ($N_s = 50, 150, 300, 400, 500, 800$ and 1000 data points), we generate N_s pairs (x_α, n_α), to which we fit with polynomials of order $q = 0, 1 \ldots 5$. Figures 5, 6, and 7 provide all diagnostic results, excepting cases when the fit involved more parameters than points.

Choosing a good fit and estimating its error: An optimal fit for a given sample minimizes I_q (top panel in Figure 5), a direct measure of the rms error of the fit. Since it measures the *exact* rms error, in practice I_q cannot be calculated; however, in our experience with this and other sample problems, this rms error can be obtained by comparing different fit orders ($J_q \approx I_q$; see the top panel of Figure 5) Thus, a comparison of different fit orders (J_q) provides both a mechanism to select the optimal fit order q and to estimate the rms error associated with it.

Alternatively, a good fit can be chosen by its performance on a *blind test* (i.e., χ^2_b): in our experience, the fit order that minimizes χ^2_b roughly agrees with the order that minimizes J_q (compare the top and bottom panels of Figure 5). Given a good fit, its rms error can be estimated from the on-sample differences between the fit and the data (σ_q; compare the top and center panels of Figure 5).

Both methods for choosing the best fit order q should and do agree. However, the two methods for estimating the rms fit error need not, particularly if the data sample does not representatively sample the space (e.g., if the sample over-represents systems with low n). Because strong biases can be easily introduced into σ_q through uneven sampling, and because in practice our population synthesis sample is highly irregular (see Figure 1), we will exclusively empoy J_q for the purposes of rms error estimation henceforth.

Interpreting the statistical diagnostics: We have chosen a trial function μ_1 which is easily fit with low orders of our basis functions; for this reason, the "statistical" diagnostics χ^2_q and L_q rapidly converge to their expected values given a perfect fit (1 and 0, respectively; see top panel of Figure 6). More generally, however, low-order basis functions need not accurately describe the function we seek to fit. In these more common cases (i.e., obtained from μ_3 and μ_5) "statistical" diagnostics deviate strongly from the values expected for a perfect fit; their absolute values reveal little to nothing about fit quality.

However, we can compare the performance of our fit in a blind test to its behavior on the original sample with the *ratio* $\chi^2_{q,b}/\chi^2_q$ (bottom panel, Figure 6). This ratio will be nearly 1 for a fit that has the same qualitative behavior on the original data set and in a blind test; however, when our fit has spurious oscillations between data points due to overfitting, this ratio will become large.

Reliable constraint prediction: As demonstrated in Figure 7 for the rather well-resolved case of μ_1, the constraint-satisfying volume $V(C, \hat{M})$ quite accurately predicts $V(C, M)$: for most well-resolved orders and most subintervals involving sufficiently large values of μ (i.e., well-sampled results), more than 90% of the predicted constraint-satisfying points do indeed satisfy the constraints ($r_+ > 0.9$) and more than 90% of the constraint-satisfying points are inside our predicted constraint-satisfying region (r_-).

When we broaden the constraint-satisfying interval by the observed rms tolerance J_q, we can with considerable confidence ($r_+ > 97\%$) claim that all constraint-satisfying points lie within $V(\tilde{C}(J_q), \hat{M})$; see Figure 8. However, by broadening our interval, we have substantially increased the number of *superfluous* population synthesis models included in the predicted volume.

Similar results arise for other well-sampled test functions. Thus, so long as the constraint interval lies within a consistently well-sampled region, we can with considerable confidence overestimate the constraint-satisfying volume by $V(\tilde{C}, \hat{M})$. In other words, we can conservatively estimate the constraint-satisfying volume by (i) constructing a "broadened" constraint interval which has been widened by our estimate of the rms systematic errors associated with the fitting process and then (ii) selecting all points which, according to our fit, produce

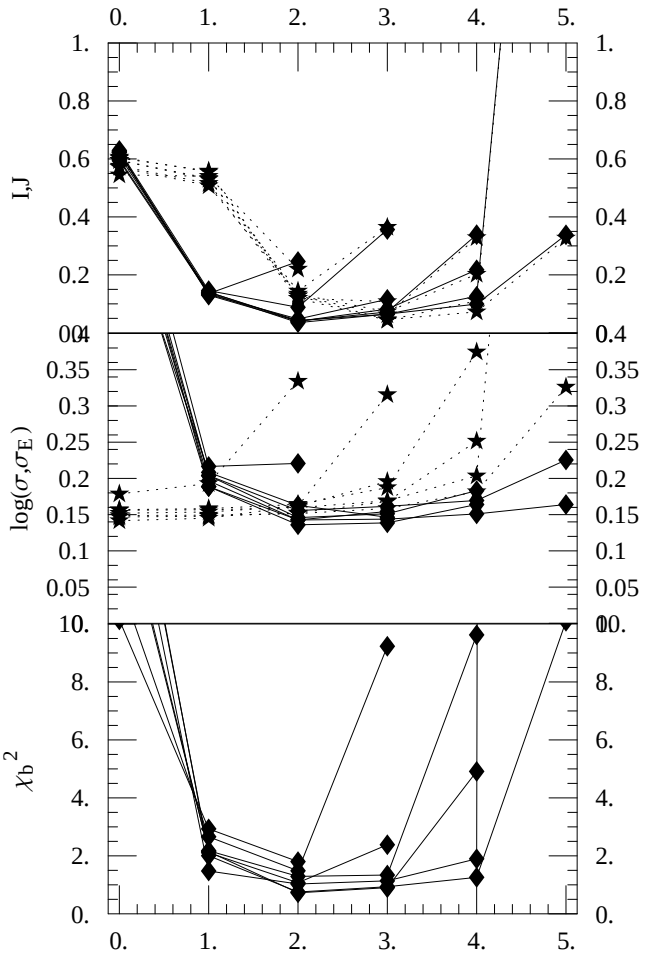

FIG. 5.— Top panel: I_q (solid) and J_q (dashed); center panel: σ_q and $\sigma_{q,E}$; bottom panel: $\chi^2_{q,b}$; all shown versus fit order q. These quantities have been evaluated for weighted least-squares fits applied to random samplings of μ_1 with N, versus the order of the fit used.

rates within the constraint-satisfying interval.

4. FITTING: POPULATION SYNTHESIS PREDICTIONS

Using our large archive of population synthesis results for event rates and relative uncertainties (§ 2), we perform a weighted least-squares fit of polynomial-like basis functions (described in Appendix A) over our seven-dimensinal space, for each of the event rates of interest: BHBH(m), BHNS(m), NSNS(m), NSNS(vw), WDNS(e), WDNS(m), SNIb/c, and SNII. Following the guidelines developed through study of toy models (§ 3), we evaluate the fit quality of several different polynomial orders to our data. To minimize the possibility of using more parameters than allowed by the number of our data points,

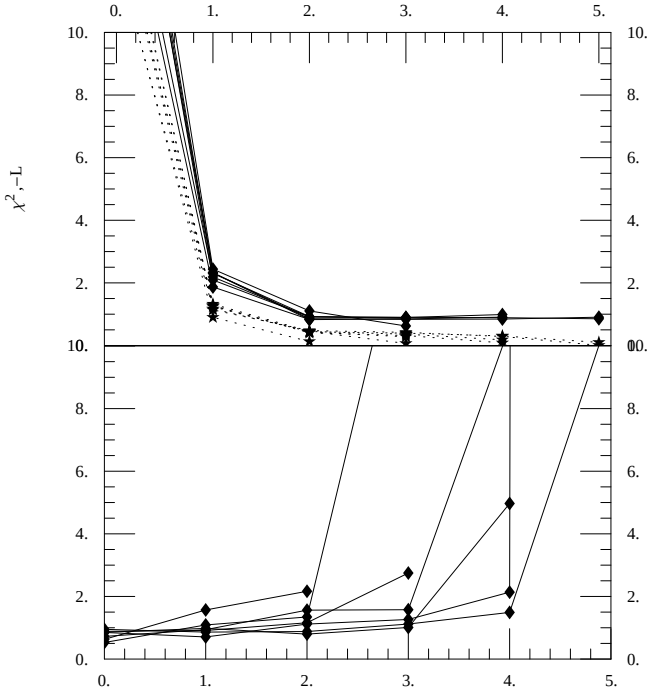

FIG. 6.— Top panel: χ^2_q (solid) and $(-L_q)$ (dotted); bottom panel: $\chi^2_{q,b}/\chi^2_q$.

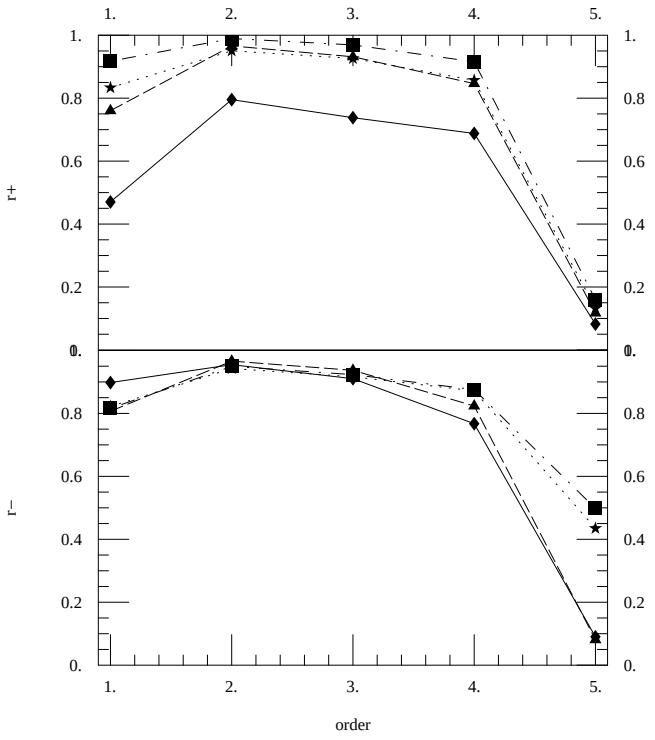

FIG. 7.— For the four constraint subintervals $C_{4,z=1...4}$, plots of $r_+(C_{q,4,z})$ (top) and $r_-(C_{q,4,z})$ versus fitting order q, based on fits applied to $N = 800$ samples of μ_1. The solid curve is associated with the lowest constraint interval ($z = 1$), followed by dotted, dot-dashed, and dashed, respectively.

we choose as "best" fit that order which minimizes the relative difference between fit orders J_q (see below).

Table 1 summarizes properties of the least-squares fits we applied to our archived population synthesis results. The first column provides a brief label for the fit, as described in greater detail in the text. The next two columns summarize the amount of available information contained in our population synthesis archive: N is the number of population synthesis archives with unbiased data (i.e., where all plausible progenitors for the target event have been included), whereas N_+ is a smaller number of archives with unbiased data containing one or more events (i.e., for which an estimate of the rate, rather than merely an upper bound, is possible). A block of five columns describes properties and diagnostics of a weighted least-squares fit applied to our data. The first two columns, q and n_{dof}, merely indicate the polynomial order and number of degrees of freedom involved in our fit; in all the cases shown here, the optimal polynomial order produces far fewer degrees of freedom than points. The next three columns provide critical diagnostics of our fit: χ^2_q, σ_q, and J_q. The first two compare our fit with the raw *sampled* data – labelled by paramter points x_k; by the specific number n_k of results seen in the corresponding population synthesis simulation; and by R_k our estimated result [Eqs. (**??**]. The third measures the rms difference between the fit used an the next-lowest-order fit; as extensively addressed in § 3, this quantity is believed to provide the best estimate for the global rms point error in the fit.

$$\chi^2_q \equiv \frac{1}{N_s - N_p} \sum [\log_{10}(R_f(x_k)/R_k)]^2 n_k (\ln 10)^2 \quad (21)$$

$$J_q \equiv \left[\int d^7x \, [\log_{10}(R_q/R_{q-1})]^2 \right]^{1/2} \quad (22)$$

$$\sigma^2 \equiv \frac{1}{N_s} \sum_{k=1}^{N_s} [\log_{10}(R_f(x_k)/R_k)]^2 \quad (23)$$

$$\sigma_R^2 \equiv \frac{1}{N} \sum_k \log_{10}(R_f(y_k))^2 - \left(\frac{1}{N} \sum_q \log_{10}(R_f(y_q)) \right)^2 \quad (24)$$

Our fits only roughly approximate the underlying functions. Concretely, for some model our fit may predict the population synthesis code would produce a BH-BH merger rate of $\mathcal{R}_o$; however, given the substantial systematic error estimated for this fit – rms uncertainty $J_q \approx 0.33$ – we can be fairly certain only that the rate lies between $10^{-2\times 0.33}\mathcal{R}_o \approx 0.22\mathcal{R}_o$ and $5\mathcal{R}_o$, based on a two-sigma confidence interval. However, these uncertainties remain considerably less than (i) the range over which these functions vary, as measured by the last column in Table 1; and more critically less than (ii) experimental constraints on these rates, where available.

Most fits have quite large χ^2 ($\gg 1$) and σ, because our low-order polynomial basis functions are not sufficiently general models to match the intrinsic variation in

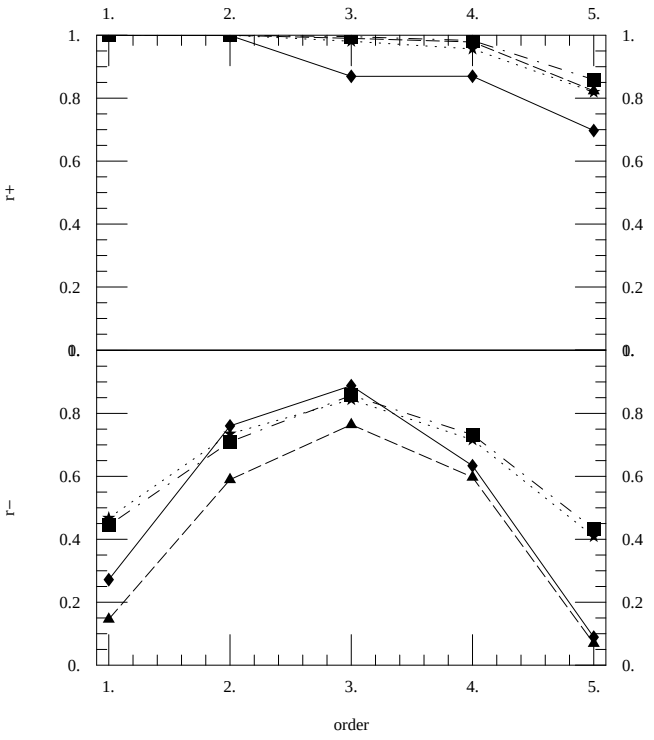

FIG. 8.— For the four constraint subintervals $C_{4,z=1\ldots4}$, plots of the fraction of constraint-satisfying points within a tolerance-broadened interval $[r_+(\tilde{C_{q,4,z}}(J_q))$; see Eq. 13 for details]. versus fitting order q, based on fits applied to $N = 800$ samples of μ_1. The solid curve is associated with the lowest constraint interval ($z = 1$), followed by dotted, dot-dashed, and dashed, respectively.

population synthesis results.[2]

4.1. *Sample Completeness*

Some population synthesis archives which provide unbiased data for event rate estimates have zero events (i.e., the number of population synthesis models N_s is more than the number of models with at least one observed event N_{s+}, in Table 1). Because of the enormous range of population synthesis sample sizes N used, sometimes zero events for rare events should be expected [i.e., merging BH-BH binaries: BHBH(m)]. On the other hand, a significant set of large-N population synthesis archives with $n = 0$ for a relatively *common* event might imply that the event has been systematically undersampled: in some region, the underlying event rate might be significantly lower than our archive collection has been able to probe. Consequently these low event rates are not represented in our database of rates to be fit.

To distinguish between the two possibilities in cases where a large number of zero-event samples ($N_s - N_{s+}$) indicate that undersampling is possible [i.e., BHBH(m) and NSNS(vw)], we scatter-plot $\log_{10} N_s$ versus $\log_{10} n$, superimposed by the line $n/N = R_{99\%}/s$, where $R_{99\%}$ is the 99% lower limit predicted on the basis of the fit (see Figure 9). If the fit is of high quality , very few systems should lie below this line; further, if our sample is fairly complete, *very* few high-N systems should lie below this line. We see no compelling indication of undersampling.[3]

4.2. *Results*

Supenovae: Given our choice of stellar mass interval probed in our simulation (i.e., $m_1 > 4M_\odot$), supernovae occur extremely frequently, providing us with superb statistics at low cost. However, our limited set of basis functions can only with difficulty reproduce the observed variation in supernovae rates: even though SN rates for models in our archive are at times known to 1% (i.e., involving 10^4 or more sampled events), or 0.004 in the log, our optimal fit differs significantly from the data, by $\sigma \approx O(0.04)$ in the log.

Nonetheless, as discussed in the forthcoming O'Shaughnessy et al. (2006c), the supernovae rate remains a striking success of the *StarTrack* population synthesis code and our normalization conventions (e.g., $\dot{M} \approx 3.5 M_\odot \mathrm{yr}^{-1}$). No matter what combination of population synthesis parameters we choose, the predicted SN rates lie well within the observational constraints found by Cappellaro et al. (1999).

WD-NS binaries: As with supernovae, white dwarf-neutron star binaries occur fairly frequently (see $\langle \log R \rangle$ in Table 1), allowing us to accumulate fairly good statistics over a broad range of population synthesis parameters. Additionally, based on the distribution of N_s versus n (i.e., as in Figure 9), our sample shows no signs of systematic incompleteness: we appear to have covered the full range. Though our polynomial fits continue to introduce systematic error, the resulting fit behaves well throughout the range.

BH-BH binaries: Double black hole binaries, in contrast, occur extremely infrequently. Nonetheless, by using special-purpose partitions, O'Shaughnessy et al. (2005a) accumulated a large ($\approx$ 500 simulations) sample with good ($n \approx 10$) statistics on a large subset of parameter space. By adjoining the results of general-purpose simulations *not* assured of good statistics, this sample has since been enlarged by a factor ≈ 5, with emphasis on the same subset of parameter space presented in O'Shaughnessy et al. (2005a) (see Figure 1). Finally, though the distribution of n versus N_s (Figure 9) suggests the lowest BH-BH merger rates may not be very well-resolved, we have no reason to suspect we have any significant under-resolved region: simulations larger than 10^6 binaries consistently produce several merger events ($n \gg 0$).

For these reasons, we are therefore not surprised to discover that the BHBH(m) fit is statistically fairly consistent given our polynomial model limitations (relatively low χ^2) but still has a fairly significant rms error (measured by σ and J, particularly in comparison with σ_R)

Results: NS-NS and BH-NS binaries: Despite (or perhaps because of) a concerted effort to accumulate good

[2] For example, when fitting supernovae rates as a function of population synthesis parameters for *single* stars, where only one of our parameters (wind strength) enters, we find the rate has a moderately complex functional form that requires high-degree polynomials to fit. We expect similarly complex behavior in the multidimensional case, and interpret the large χ^2 correspondingly.

[3] Unfortunately, the hardest simulations had to be run the longest, typically producing an over-represented collection of low-n, high-N simulations. As a result, the distribution of N and n is often biased at the largest N.

statistics targeted specifically to these classes, fits for NS-NS and BH-NS event rates are statisticallly implausible as measured by χ^2. Judging from Figure 9, the NSNS(m) and BH-NS(m) are comparatively well-resolved: longer simulations fairly consistently have a lower chance of producing $n = 0$ results. For this reason, we strongly suspect some feature of these underlying rate functions are poorly-described by our basis functions; we intend to more thoroughly test this hypothesis (with better statistics) by comparing these fits to nonparametric estimates in a future paper.

Again judging from Figure 9, the sparsely-sampled NSNS(vw) rate is likely undersampled: the longest simulations, oddly, are *more* likely than usual to have $n = 0, 1$. Though we are at present unable to directly characterize the region responsible, we suspect that certain combinations of population synthesis parameters produce an abnormally low probability for recycling pulsars in wide NS-NS binaries. This systematic undersampling is presumably responsible in large part for the unusually poor quality of the fit ($J_q \approx 0.8$, more than twice as large as any other error bar).

5. CONCLUSIONS

To develop a more comprehensive understanding of population synthesis predictions and to allow those theoretical predictions to be systematically compared with observations of the end products of high-mass single and binary stellar evolution, we have fit eight predictions from the *StarTrack* code over seven of its input parameters. These fits are available on requests sent to the first author. In a forthcoming paper, O'Shaughnessy et al. (2006c) will apply these fits along with estimates of their systematic errors to discover robust constraints on the seven parameters that enter into population synthesis. Additionally, we have demonstrated that in analagous model problems the constraint-satisfying region defined by using these fits can, under appropriate conditions, very accurately trace the underlying constraint-satisfying region. Finally, we have presented a thorough diagnostic formalism, including a large list of diagnostic quantities and tests ($I, J, \sigma, \chi^2, L, \eta, r_\pm$, etc.), which can be applied to investigate any fitting method on any population synthesis code.

We thank NCSA for providing us with resources used to perform many of the computations in this text. This work is partially supported by an NSF Gravitational Physics grant PHYS-0353111, a David and Lucile Packard Foundation Fellowship in Science and Engineering, and a Cottrell Scholar Award from the Research Corporation to VK. KB acknowledges support from the Polish Science Foundation (KBN) Grant 1P03D02228. VK is grateful for the hospitality and partial financial support of the KITP, UCSB for the period of June and July 2006.

REFERENCES

Abbott, B., Abbott, R., Adhikari, R., Ageev, A., Allen, B., Amin, R., Anderson, S. B., Anderson, W. G., Araya, M., Armandula, H., Ashley, M., Asiri, F., Aufmuth, P., Aulbert, C., Babak, S., Balasubramanian, R., & OTHERS. 2005a, Phys. Rev. D, 72, 082002 [ADS] [ADS]

Abbott, B., Abbott, R., Adhikari, R., Ageev, A., Allen, B., Amin, R., Anderson, S. B., Anderson, W. G., Araya, M., Armandula, H., Asiri, F., Aufmuth, P., Aulbert, C., Babak, S., Balasubramanian, R., Ballmer, S., Barish, B. C., Barker, D., Barker-Patton, C., & OTHERS. 2005b, Phys. Rev. D, 72, 082002

Arzoumanian, Z., Cordes, J. M., & Wasserman, I. 1999, ApJ, 520, 696 [ADS]

Belczynski, K., Kalogera, V., & Bulik, T. 2002, ApJ, 572, 407 [ADS]

Belczynski, K., Kalogera, V., Rasio, F., Taam, R., Zezas, A., Maccarone, T., & Ivanova, N. 2006, arxiv eprint [URL]

Cappellaro, E., Evans, R., & Turatto, M. 1999, A&A, 351, 459 [ADS]

Fryer, C., Burrows, A., & Benz, W. 1998, ApJ, 496, 333 [ADS]

Hobbs, G., Lorimer, D. R., Lyne, A. G., & Kramer, M. 2005, MNRAS, 360, 974 [ADS]

Hurley, J. R., Tout, C. A., & Pols, O. R. 2002, MNRAS, 329, 897 [ADS]

Kalogera, V., Kim, C., Lorimer, D. R., Ihm, M., & Belczynski, K. 2005, in ASP Conf. Ser. 328: Binary Radio Pulsars, ed. I. Stairs, 261–+

Kim, C., Kalogera, V., & Lorimer, D. R. 2003, ApJ, 584, 985 [ADS]

Kim, C., Kalogera, V., Lorimer, D. R., & White, T. 2004, ApJ, 616, 1109 [ADS]

Lacey, C. G. & Fall, S. M. 1985, Astrophysical Journal, 290, 154

Lorimer, D. 2005 [URL]

O'Shaughnessy, R., Belczynski, C., Perna, R., Bulik, T., Kalogera, V., & Lamb, D. 2006a, in preparation

—. 2006b, in preparation

O'Shaughnessy, R., Kalogera, V., & Belczynski, K. 2005a, ApJ, 620, 385

O'Shaughnessy, R., Kim, C., Fragos, T., Kalogera, V., & Belczynski, K. 2005b, ApJ, 633, 1076 [ADS] [ADS]

O'Shaughnessy, R., Kim, C., Kalogera, V., & Belczynski, K. 2006c, In preparation

Portegies Zwart, S. F. & Verbunt, F. 1996, A&A, 309, 179 [ADS]

Rana. 1991, Annual reviews of astronomy and astrophysics, 29, 129 [URL]

APPENDIX

POLYNOMIAL BASIS FUNCTIONS

To parameterize supernovae kick distributions, *StarTrack* employs three parameters, σ_1, σ_2, and s, which represent the superposition of two Maxwellian kick distributions with probabilities s and $1-s$. The physical predictions associated with (σ_1, σ_2, s) are therefore identical to those of $(\sigma_2, \sigma_1, 1-s)$. To improve the physical significance of our fit, we have chosen to employ basis polynomials which *enforce* this requirement to all orders.

Specifically, rather than allow for homogeneous basis polynomials in these parameters, we use the following, for arbitrary p and q:

$$\sigma_1^p s^q + \sigma_2^p (1-s)^q \tag{A1}$$

$$2\sigma_1^p \sigma_2^p s(1-s) + \sigma_1^{2p} s^2 + \sigma_2^{2p} (1-s)^2 \tag{A2}$$

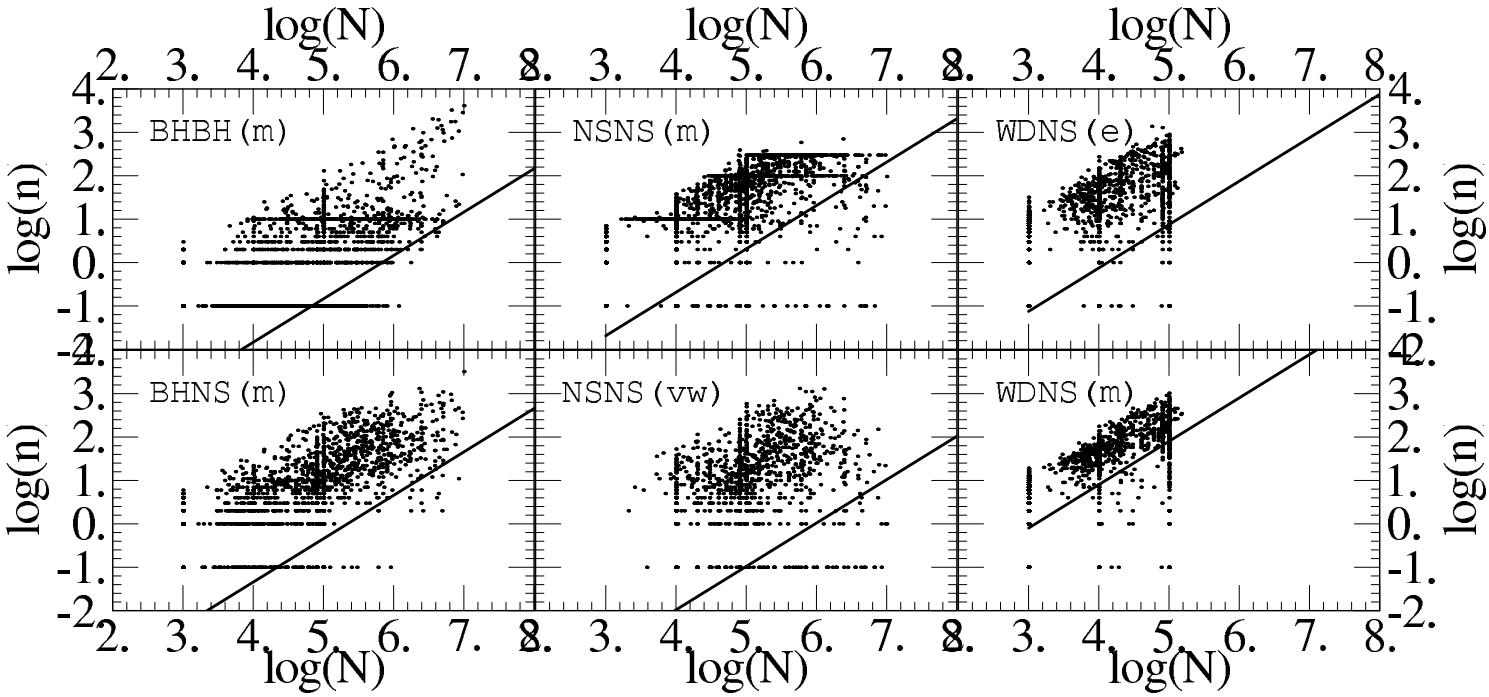

FIG. 9.— Log-log scatter plot of the number of events seen in a sample against the population synthesis sample size. Also shown is a diagonal line corresponding to the 99% lower bound on the fit-predicted rate distribution, translated from physical rate into expected number of events seen per unit population synthesis sample size. For completeness, models with zero events are shown with $\log_{10} n$ as -1.

TABLE 1. STATISTICS OF POPULATION SYNTHESIS SIMULATIONS FOR MERGER RATES

Type[a]	N_s^b	$N_{s,+}^c$	n_{dof}^d	q^e	$\chi^{2,f}$	σ^g	J_q^h $\log(\mathrm{yr}^{-1})$	$\eta(10)^g$	$\min \log \mathcal{R}_f(S)^h$ $\log(\mathrm{yr}^{-1})$	$\max \log \mathcal{R}_f(S)^i$	$\langle \log R \rangle_S^j$ $\log(\mathrm{yr}^{-1})$	σ_R^k $\log(\mathrm{yr}^{-1})$
BH-BH(m)	2930	1351	141	3	7.0	0.56	0.33	4%	-4.2	-7.5	-6.0	0.61
BH-NS(m)	2533	1627	141	3	12.8	0.41	0.28	1%	-3.8	-8.0	-5.5	0.58
NS-NS(m)	2803	2272	141	3	15.1	0.36	0.21	7%	-3.8	-6.8	-4.9	0.53
NS-NS(vw)	1325	1051	46	2	11.2	0.38	0.39	10%	-3.5	-8.6	-5.6	0.74
WD-NS(e)	1770	1565	141	3	11.6	0.33	0.19	10%	-3.0	-6.5	-4.1	0.54
WD-NS(m)	1770	1642	141	3	14.2	0.35	0.18	12%	-3.0	-5.0	-3.8	0.28
SN Ib/c	1482	1482	141	3	7.8	0.07	0.06	n/a	-1.8	-2.4	-2.1	0.10
SN II	1482	1482	141	3	6.0	0.04	0.02	n/a	-2.0	-3.3	-2.4	0.16

[1]Type of event of interest; types are defined in the text.
[2]Number of population synthesis simulations for which data regarding this rate exists, unbiased by (for example) limitations on binary parameters.
[3]Number of population synthesis simulations for which good data exists and *one or more* events of interest occur.
[4]Degree of the polynomial fitted to seven-dimensional population synthesis simulation data; see Sec. **??**.
[5]See Eq. 21.
[6]The rms deviation of data points from the fit; see Eq. 23.
[7]The fraction of those simulations with *no* events for which the rate estimate, combined with sample size, predicts more than 10 events should have been observed.
[8]Given 2000 random sample points S, the minimum value R_f takes on the sample. When bounded to a reasonable range, a crude estimate of global fit quality.
[9]Similarly for the maximum.
[9]Mean value of the fit on S. An estimate of the a priori plausible event rate produced by population synthesis.
[10]Standard deviation of the fit on S. An estimate of the width of the distribution of plausible event rates.

Because s must enters in a heterogeneous manner to preserve our desired symmetry, these basis polynomials are of a fixed order in all kick parameters. For the purposes of order counting when constructing fits, the first polynomial is denoted order $p + q$ and the second order p.